



**caBIG** *cancer Biomedical  
Informatics Grid*



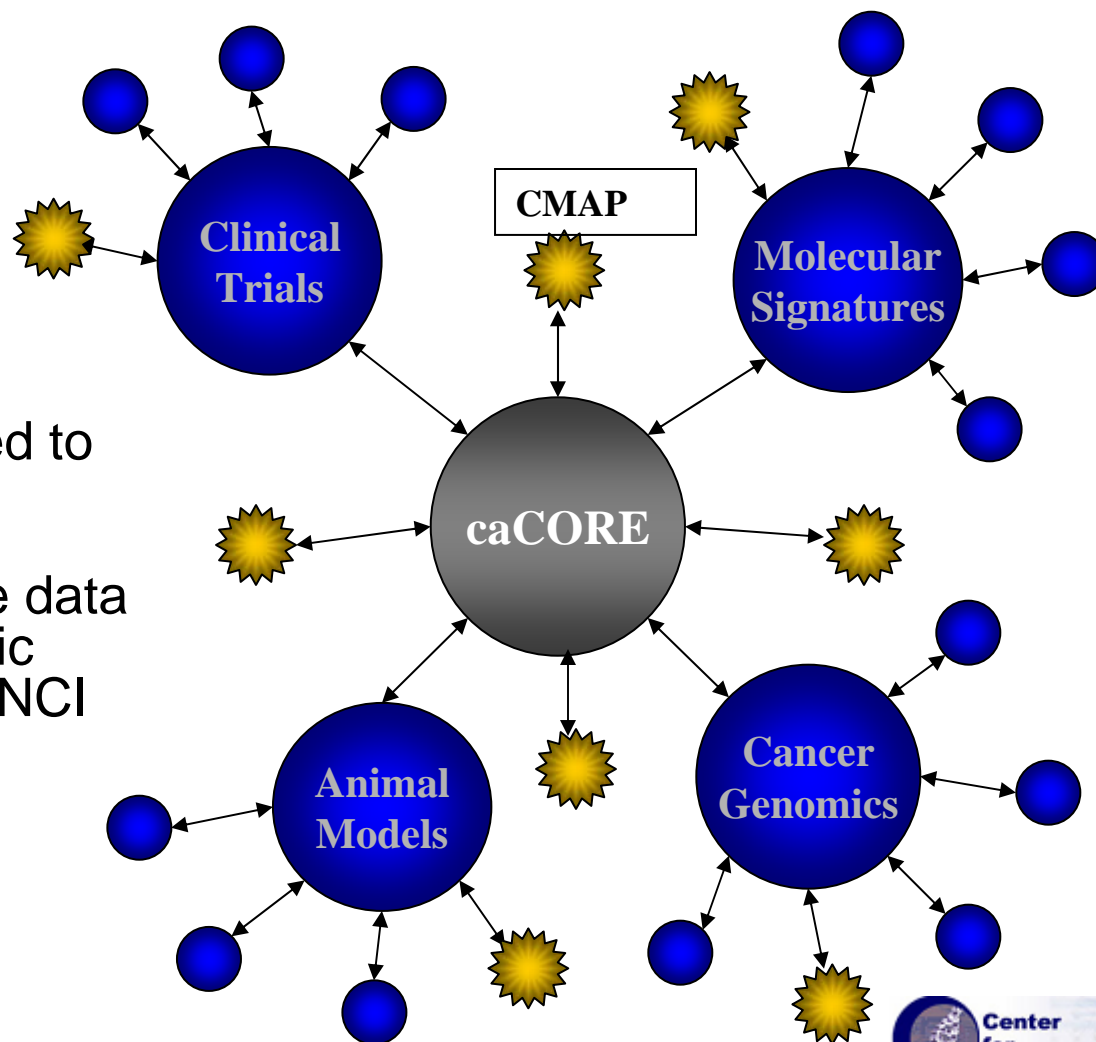
## **cancer Bioinformatics Infrastructure Objects (caBIO)**

***NCICB Informatics Supporting Translational Research***

***Himanso Sahni***

## The NCICB Cancer “Core”

The NCICB is dedicated to building common architecture, tools, and standards that facilitate data integration and scientific collaboration between NCI research initiatives



# caCORE

- ▶ cancer Common Ontologic Representation Environment
- ▶ caCORE is the technology stack that facilitates data integration across multiple scientific disciplines

Enterprise Vocabulary Services (EVS)

Cancer Data Standards Repository (caDSR)

Cancer Bioinformatics Infrastructure Objects (caBIO)

# What is caBIO?

## ▶ Question

- Is it a bio-informatics Infrastructure?
- Is it a data source?
- Is it an open source software project?
- Is it a beach front resort hotel in Spain?

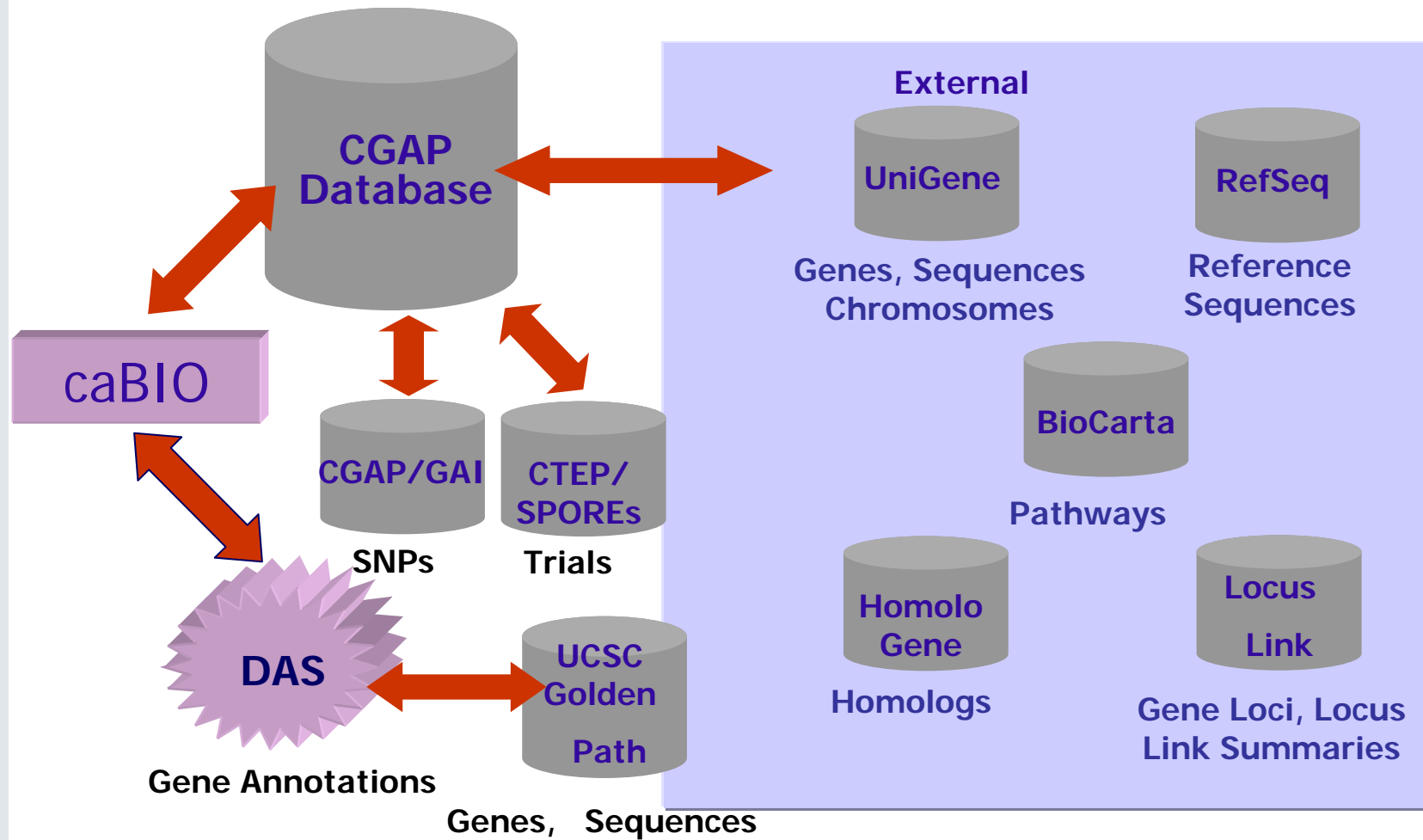
## ▶ Answer

- All of the above 😊

# caBIO Infrastructure serves caCORE technologies

- ▶ Provides standard object models and a uniform programmatic interface access (Java, web services (SOAP), and HTTP) to the entire caCORE technologies
  - caBIO data sources
  - caDSR (Metadata Repository)
  - Enterprise Vocabulary Services
    - NCI Metaphrase Vocabulary servers
    - NCI Thesaurus (Descriptive Logic) servers
  - Cancer Models Database (caMOD)

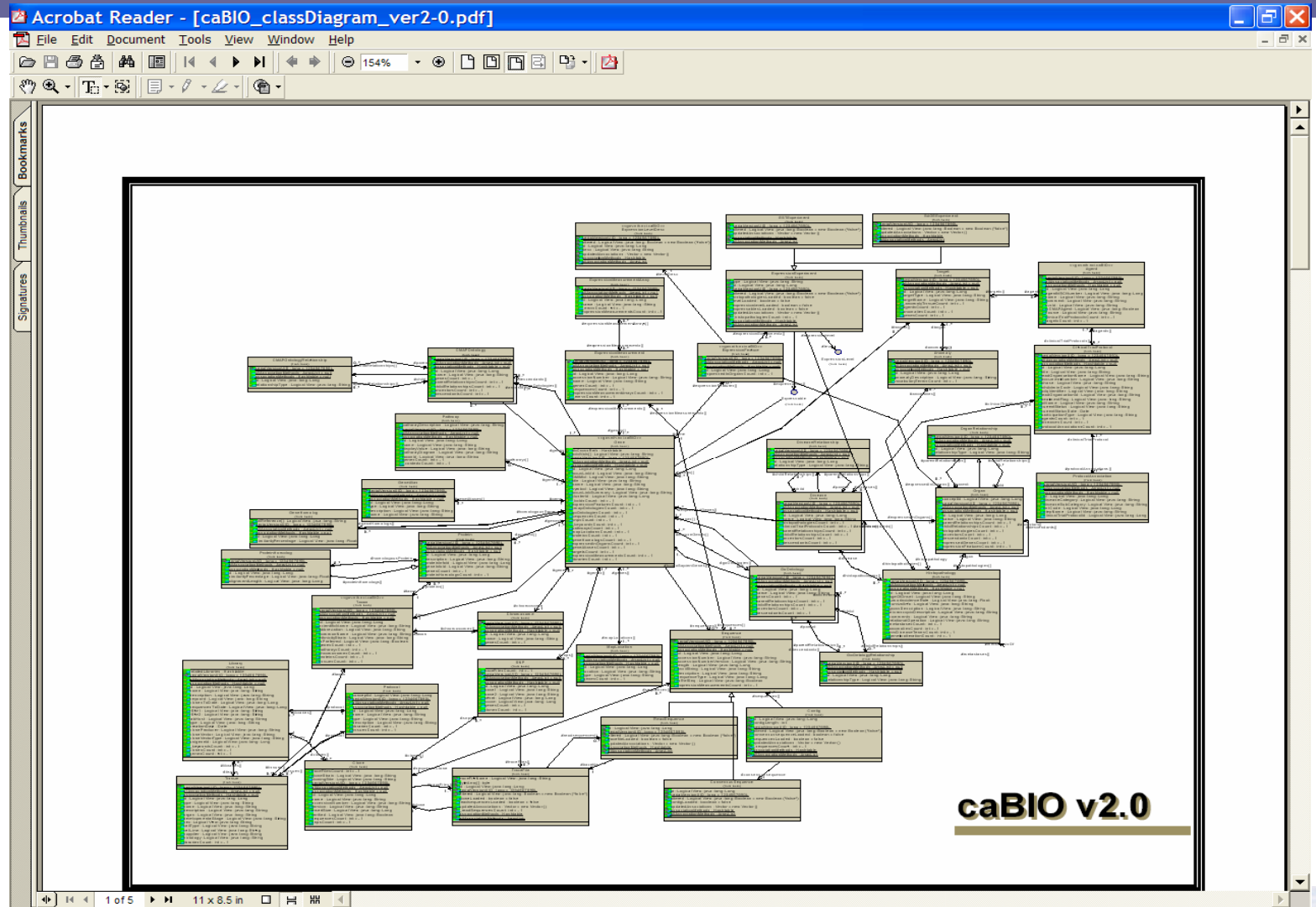
# caBIO Data Sources



## caBIO Software Project

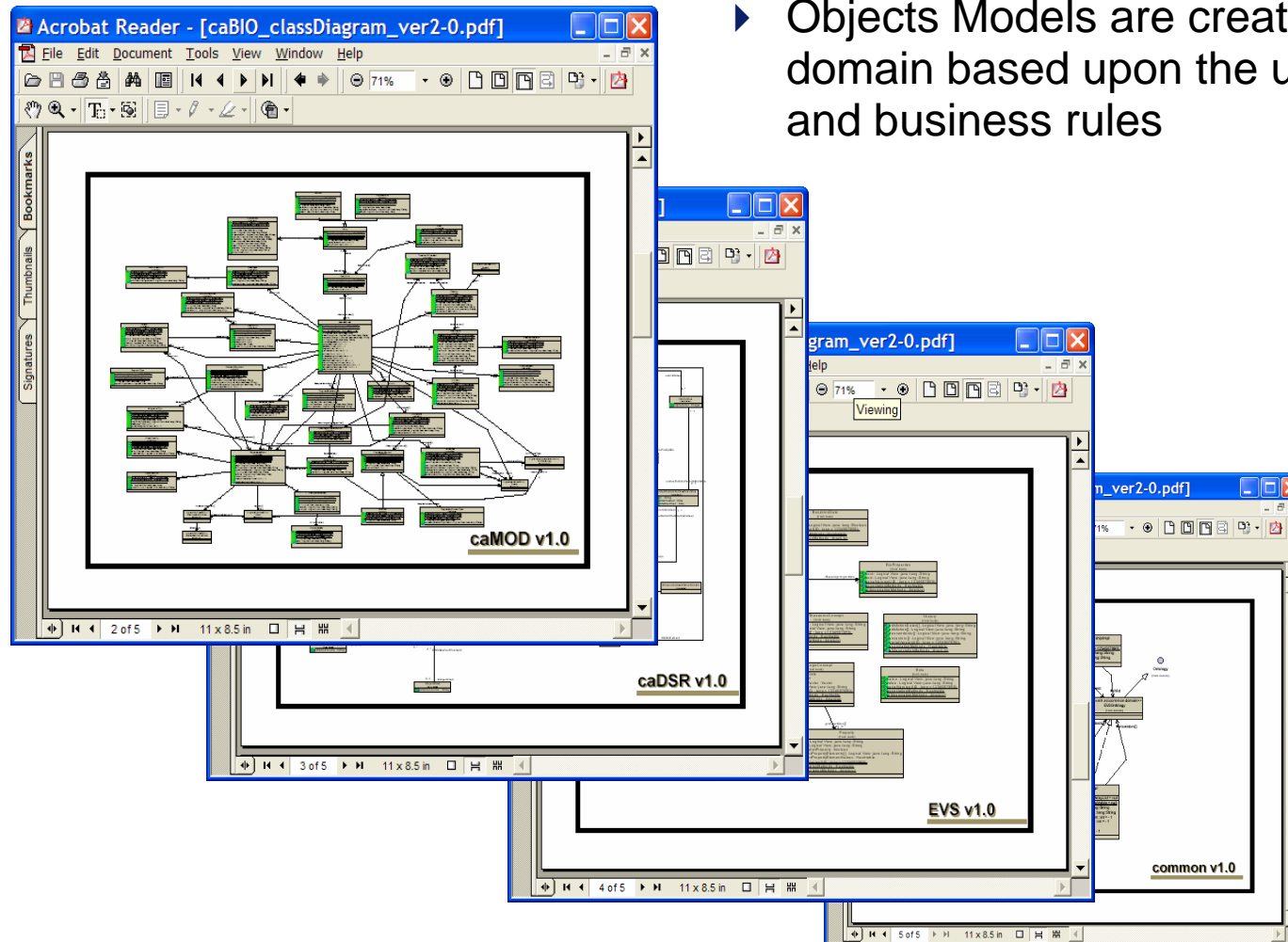
- ▶ Is an “open source” project built upon open source technologies.
- ▶ It provides an abstraction layer that allows developers to access genomic, systems biology, clinical and pre-clinical, biomedical metadata and a wide variety of medical vocabularies.
- ▶ It uses a standardized tool set without concerns for implementation details and data management.
- ▶ <http://ncicb.nci.nih.gov/core/caBIO>

# caBIO Object Model



## Model Extensions in caBIO

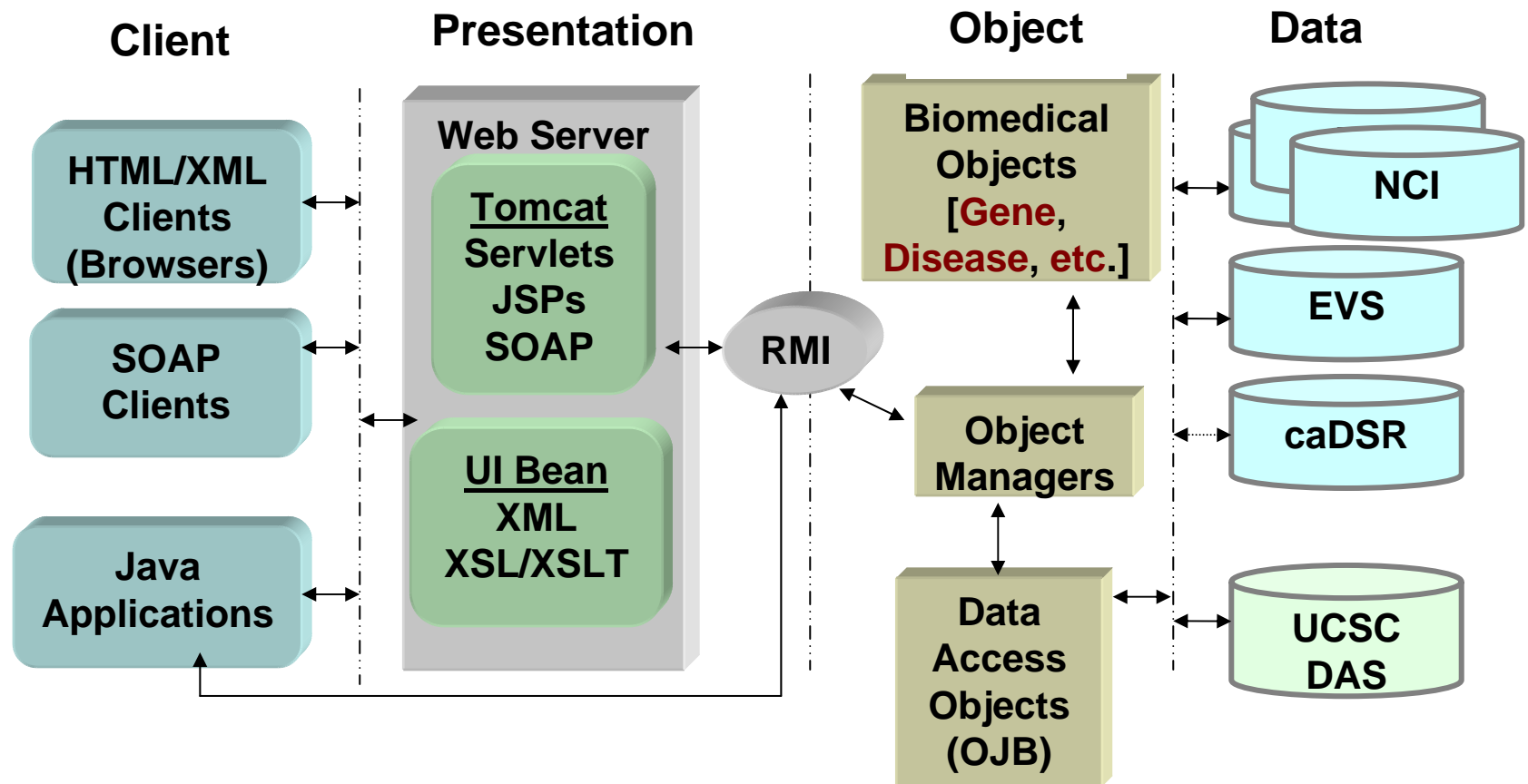
- Objects Models are created for each domain based upon the use cases and business rules



## caBIO Architecture

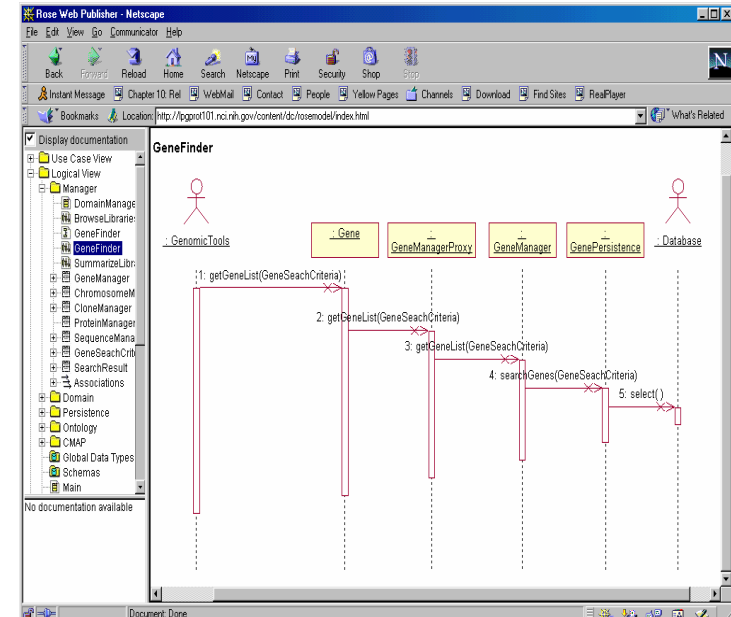
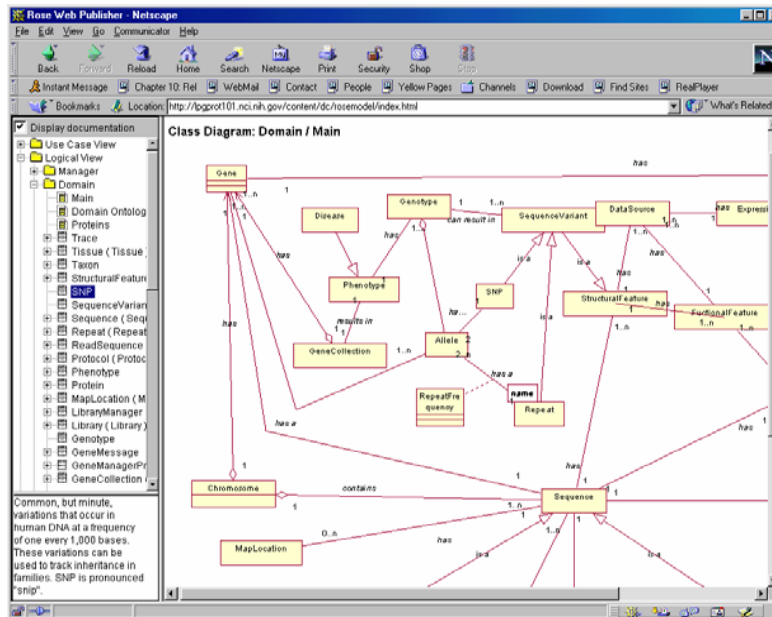
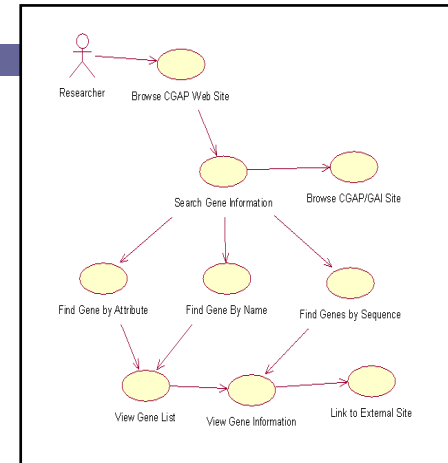
- ▶ caBIO is designed using a J2EE architecture with client interfaces, server components, back-end objects and data sources
- ▶ Stand alone or server side java applications communicate with back-end objects via Java RMI
- ▶ Non-Java based applications can communicate via SOAP or HTTP API
- ▶ Back-end objects are mapped with data sources via Apache's ObjectRelationalBridge (OBJ), and other sources (URLs, flat files)
- ▶ caBIO web services may be advertised to facilitate information sharing

# Architecture



## Development Process: UML

- ▶ Use Cases
- ▶ Class Diagrams
- ▶ Sequence Diagrams
- ▶ Iterative Development



## caBIO Development Tools

- ▶ Today
  - Rose for UML
  - Quava for code generation
- ▶ End of 2004
  - Poseidon for UML
  - Axgen for code generation
  - Low/no cost, better support for open standards

# caBIO APIs

## ► Java

- Query/retrieve biomedical objects directly via RMI

## ► SOAP

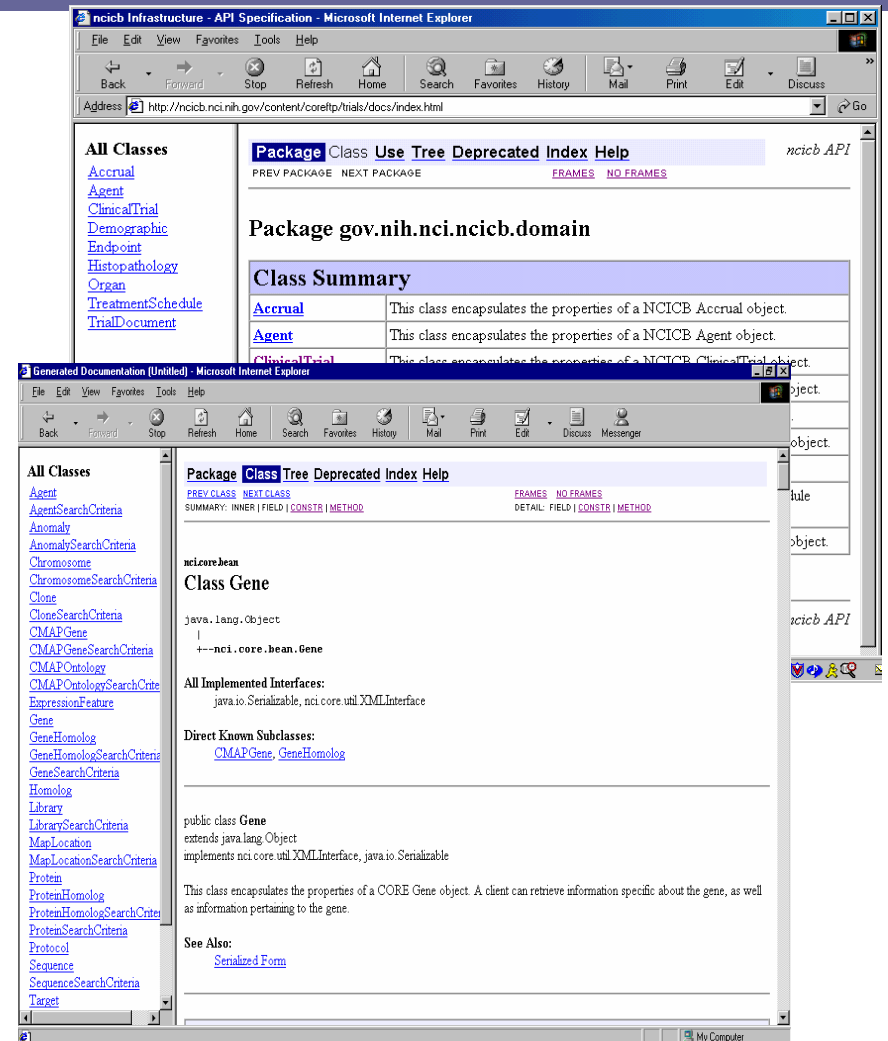
- SOAP client in any language/environment can send request to NCICB server for object data
- SOAP-XML envelope and payload returned

## ► HTTP-XML

- Properly formed URLs in any web browser/client can retrieve XML-formatted object data directly

## ► caBIO Perl

- Coming soon!



# Java Packages

- ▶ **gov.nih.nci.caBIO.bean**
  - Contains domain objects to access genomic and biomedical components
- ▶ **gov.nih.nci.caBIO.util.das**
  - Primary interface to the UCSC DAS
  - Uses JAXB to convert DAS DTDs to objects
- ▶ **gov.nih.nci.EVS.bean**
  - Provides programmatic access to the the NCI's Enterprise Vocabulary System (EVS)
- ▶ **gov.nih.nci.caDSR.bean**
  - Provides programmatic access to the the NCI's Metadata repository (caDSR)
- ▶ **gov.nih.nci.caMOD.bean**
  - Provides programmatic access to the the NCI's Model Organism Database (caMOD)
- ▶ **gov.nih.nci.common**
  - Contains common interfaces, abstract classes and utilities

## Java API

- ▶ Domain objects have companion *SearchCriteria* objects

```
Gene myGene = new Gene();  
GeneSearchCriteria criteria = new GeneSearchCriteria();  
criteria.setSymbol("pTEN");  
  
SearchResult result = myGene.search(criteria);  
Gene[] genes = (Gene[]) result.getResultSet();
```

- ▶ caBIO supports nested *SearchCriteria*
  - *SearchCriteria* from one object type can be fed as parameters into *SearchCriteria* of another type.
- ▶ Complex queries without any SQL

## caBIO Usage

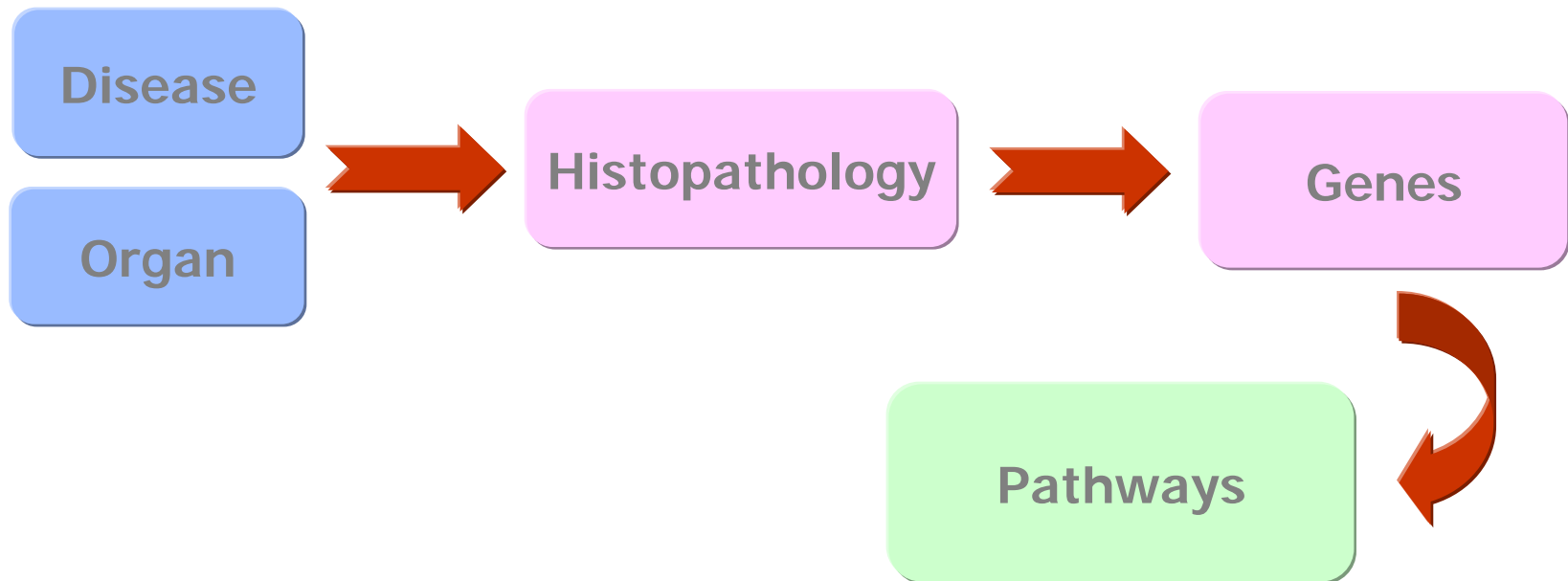
- ▶ Facilitates solving Complex Queries such as:

Find me the **Pathways**, with **Genes** that are **expressed** in **tissues** with a particular **Histopathology** that includes a particular **Organ** and a particular **Disease**.

## Traverse Relationships in Model

Find me the **Pathways**, with **Genes** that are **expressed** in **Tissues** with a particular **Histopathology** that includes a particular **Organ** and a particular **Disease**.

### INPUT



### OUTPUT

## findPathway

- Input disease, organ; create *SearchCriteria* Objects:

```
public Pathway[] findPathway(String disease, String organ) {  
    DiseaseSearchCriteria diseaseCriteria =  
        new DiseaseSearchCriteria();  
    OrganSearchCriteria organCriteria =  
        new OrganSearchCriteria();  
    HistopathologySearchCriteria histoCriteria =  
        new HistopathologySearchCriteria();  
    GeneSearchCriteria geneCriteria =  
        new GeneSearchCriteria();  
    PathwaySearchCriteria pathCriteria =  
        new PathwaySearchCriteria();  
}
```

## findPathway

- ▶ Nest the *SearchCriteria*, then do the search:

```
di seaseCri teri a. setName(di sease);  
organCri teri a. setName(organ);  
  
hi stoCri teri a. putSearchCri teri a(di seaseCri teri a, Cri teri aEl ement. AND);  
  
hi stoCri teri a. putSearchCri teri a(organCri teri a, Cri teri aEl ement. AND);  
  
geneCri teri a. putSearchCri teri a(hi stoCri teri a, Cri teri aEl ement. AND);  
  
pathCri teri a. putSearchCri teri a(geneCri teri a, Cri teri aEl ement. AND);  
  
Pathway myPathway = new Pathway();  
  
return myPathway. searchPathways(pathCri teri a);  
}
```

# findPathways: Query Results

SVGQuery - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Size Print

# CMAP

Cancer Molecular Analysis Project

Molecular Profiles Molecular Targets Molecular Targeted Agents Trials

**Current Context**

Tissue:  
**brain**

Histology Type:  
**astrocytoma**

Histology Subtype:  
**glioblastoma\_multiforme**

[Change Context](#)

[About CMAP](#)

**Quick Links:**

- [NCI Home](#)
- [NCICB Home](#)

**NATIONAL CANCER INSTITUTE**

**Pathway information provided by BioCarta**  
(See [Terms and Conditions of use](#))  
For information on sources of Pathway diagrams, see BioCarta Pathways [HowTo](#)

**A**

- Acetylation and Deacetylation of RelA in Nucleus
- Activation of cAMP-dependent Protein Kinase, PKA
- Activation of PKC through G-Protein Coupled Receptors
- Adhesion Molecules on Lymphocyte
- AhR Signal Transduction Pathway
- AKT Signaling Pathway
- Alpha-synuclein and Parkin-mediated Proteolysis in Parkinson's Disease
- Alternative Complement Pathway
- Anthrax Toxin Mechanism of Action
- Antigen Dependent B Cell Activation
- Apoptotic Signaling in Response to DNA Damage
- ATM Signaling Pathway
- Attenuation of GPCR Signaling

**B**

- B Cell Receptor Complex
- B Lymphocyte Cell Surface Molecules
- Basic mechanism of action of PPARa, PPARb(d) and PPARg and effects on gene expression
- BCR Signaling Pathway
- Beta-Oxidation of Fatty Acids
- Bioactive Peptide Induced Signaling Pathway

Internet

# Web Services: SOAP

**Deployed Service Information**

**'urn:nci-gene-service' Service Deployment Descriptor**

Property	Details
ID	urn:nci-gene-service
Scope	Application
Provider Type	java
Provider Class	gov.nih.nci.caBIO.webservices.GeneService
Use Static Class	false
Methods	getGenes, getSequences, getGeneHomologs, getReferenceSequences, getGenomicSequences, getGoOntologies, getMapLocations, getPathways, getProteins, getSNPs, getGeneAliases
Type Mappings	
Default Mapping	
Registry Class	

<http://cabio.nci.nih.gov/soap/services/index.html>

# SOAP API

## ► Perl Example

```
use SOAP::Lite;
$s = SOAP::Lite
    ->uri (urn:nci-gene-service)
    -
    >proxy("http://cabi.o.nci.nih.gov/soap/service/rpcrouter");

my %searchCriteria=();
$searchCriteria{symbol}="pTEN";
$som=$s->getGenes(SOAP::Data->type(map =>\%searchCriteria));
$xml doc = $som->result;
```

# SOAP output with xlink

```
<?xml version="1.0" encoding="UTF-8" ?>
<nci-core>
  - <gov.nih.nci.caBIO.bean.Gene id="2221" xmlns:xlink="http://www.w3.org/1999/xlink/">
    <name>PTEN</name>
    <title>phosphatase and tensin homolog (mutated in multiple advanced cancers 1)</title>
    <dbCrossRefs>{LOCUS_LINK=5728, OMIM=601728, UNIGENE=10712}</dbCrossRefs>
    <Pathway xlink:href=
      "http://cabio.nci.nih.gov/CORE/GetXML?operation=Pathway&GeneId=2221" />
    [Additional xlink for ExpressionExperiment, Organ, Chromosome, GeneHomolog,
     Sequence, Gene Alias, Protein, SNP, and MapLocation]
  </gov.nih.nci.caBIO.bean.Gene>
  [2 Additional Genes with "PTEN" in their name]
  - <searchResult>
    <hasMore>>false</hasMore> <startsAt>1</startsAt> <endsAt>3</endsAt>
  </searchResult>
</nci-core>
```

## HTTP API

- ▶ Direct access to XML-formatted data via URLs:

`http://cabio.nci.nih.gov/servlet/GetXML?  
query=Chromosome&crit_genes_symbol=pten`

Object to query

Parameter Value

Search Parameter

[http://cabio.nci.nih.gov/servlet/GetXML?query=Chromosome&crit\\_genes\\_symbol=pten](http://cabio.nci.nih.gov/servlet/GetXML?query=Chromosome&crit_genes_symbol=pten)

# HTTP API

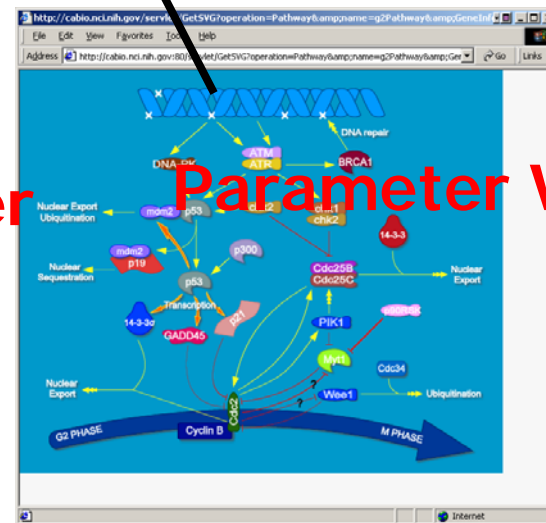
Direct access to SVG-formatted data via URLs:

<http://cabio.nci.nih.gov:80/servlet/GetSVG?operation=Pathway&name=g2Pathway&GeneInfoLocation=/servlet/GetXML?operation=Gene&ielikes=.svg>

Method

Search Parameter

Parameter Value



Powered  
by  
caBIO!

Powered by caBIO!

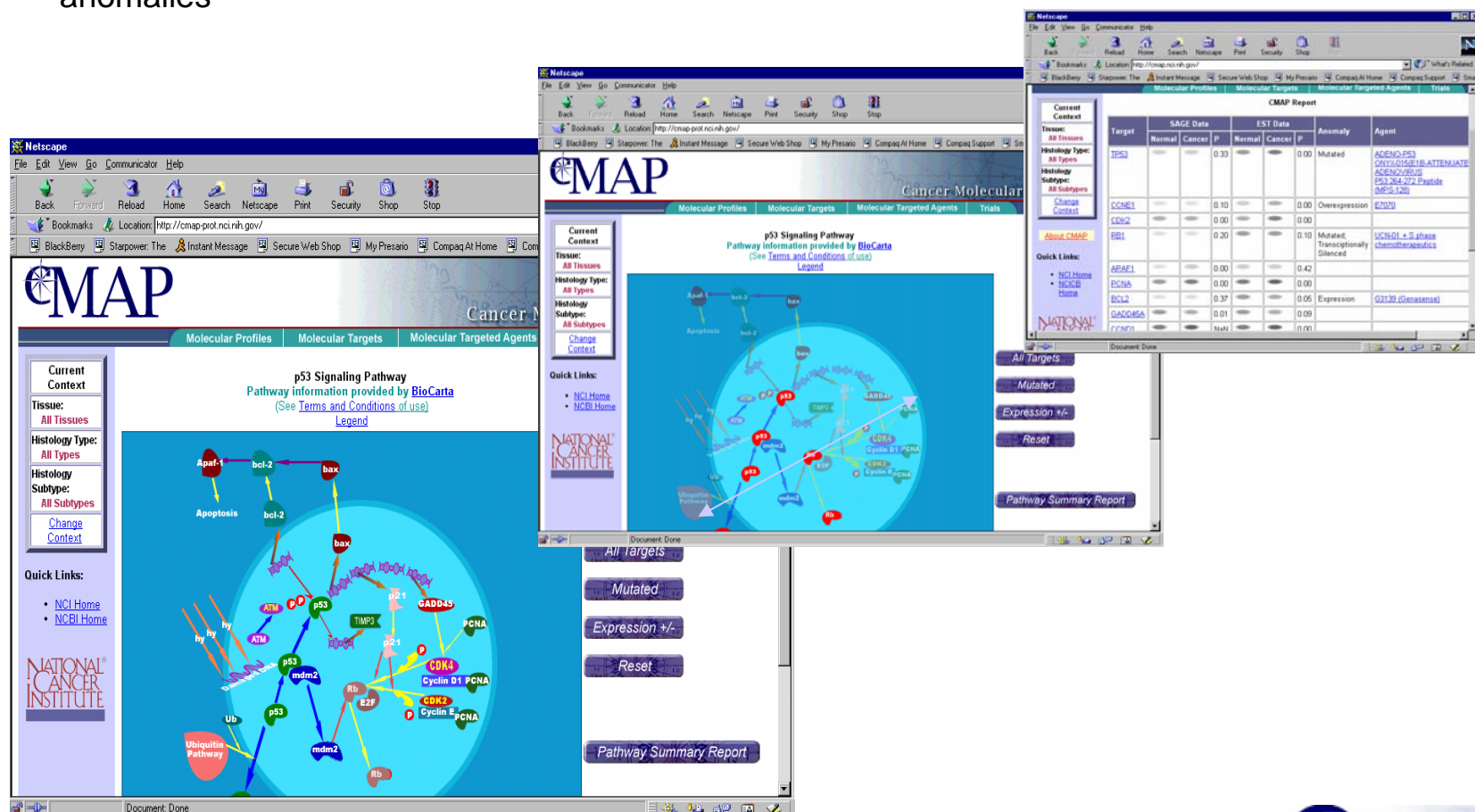
The collage features several web interfaces from the National Cancer Institute's caBIO suite:

- CMAP (Cancer Molecular Analysis Project)**: A central interface showing molecular profiles, targets, and agents, with a sidebar for context and quick links.
- MMHCC Cancer Models Database**: A database for mouse models of human cancer, showing search results and publication abstracts.
- Biogopher**: A bioinformatics tool for integrating local data with NCI infrastructure, featuring a search interface and a welcome message.
- Tree Navigator v1.1**: A tool for navigating a data tree, showing a search bar and a list of tissue types.
- Cancer Workbench (caWorkbench)**: A suite of tools for loading, visualizing, and analyzing gene expression data, shown in a separate window at the bottom left.

In the center of the collage is a graphic of the Vitruvian Man, with the text **Bio Browser** overlaid on it.

# Molecular Targets

- ▶ A collection of genes organized by pathways can be displayed facilitating the evaluation of anomalies



# What is BIOgopher?

- ▶ To biologists:
  - A spreadsheet annotation tool.
  - An ad hoc querying and reporting tool.
- ▶ To developers:
  - A GUI to the caBIO API.
  - A collection of reusable UI components.

The screenshot displays the BIOgopher web application in a Microsoft Internet Explorer browser. The address bar shows the URL <http://biogopher.nci.nih.gov/BIOgopher/index.jsp>. The page features a header with the National Cancer Institute logo and the BIOgopher logo, which includes the text 'POWERED BY caBIO!'. Below the header, there are three numbered steps: 1. optional: select local spreadsheet files to be used as data sources; 2. specify criteria for querying caBIO; 3. build report based upon query results merged with local sources (if any). The main content area contains a 'WELCOME TO BIOGOPHER!!!' message and a list of features. On the left, there is a sidebar with a 'CLICK HERE TO' section and a list of sequence accession numbers. The bottom of the page shows a table of results with columns for accession number, description, and other details.

## Two example queries

- ▶ Gene Query
  - Show me all genes and their associated pathways for all the genes that have sequences identified by the accession numbers in my spreadsheet.
- ▶ Pathway Query
  - Show me all pathways associated with genes that are expressed in tissues having a particular histology that includes brain and glioblastoma multiforme.

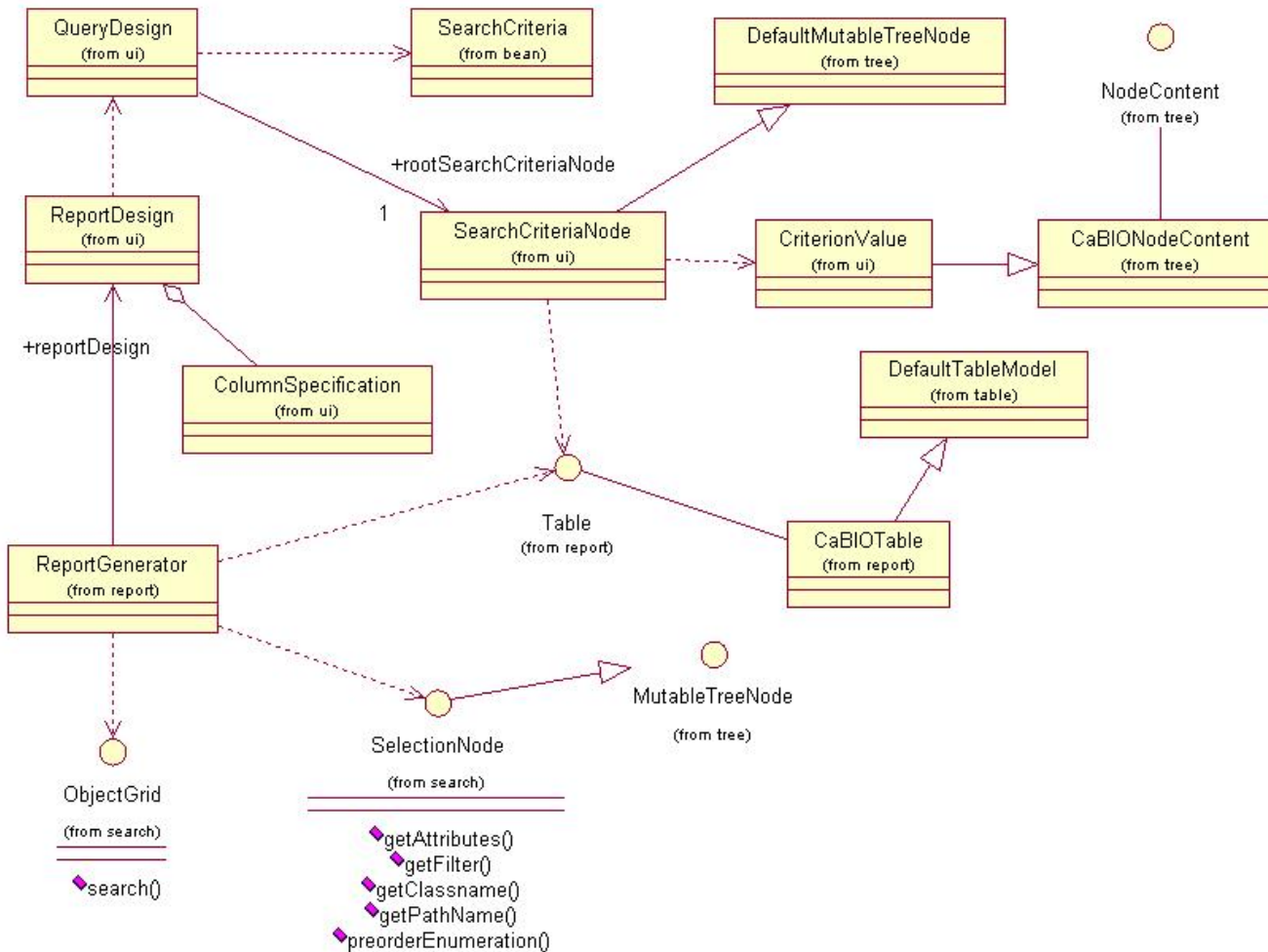
## BlOgopher Architectural Details

- ▶ Leveraged the Model-View-Controller 2 (MVC 2) architecture
  - Abstracted the presentation layer from spreadsheet manipulation, meta-data retrieval, query design, and report generation
- ▶ Utilizes caBIO's N-dimensional query builder
  - Uses an object-matrix concept to support object-mining

## For the developer

- ▶ High-level API to caBIO
- ▶ Spreadsheet parsing
- ▶ Tree manipulation
- ▶ Paging
- ▶ Metadata layer

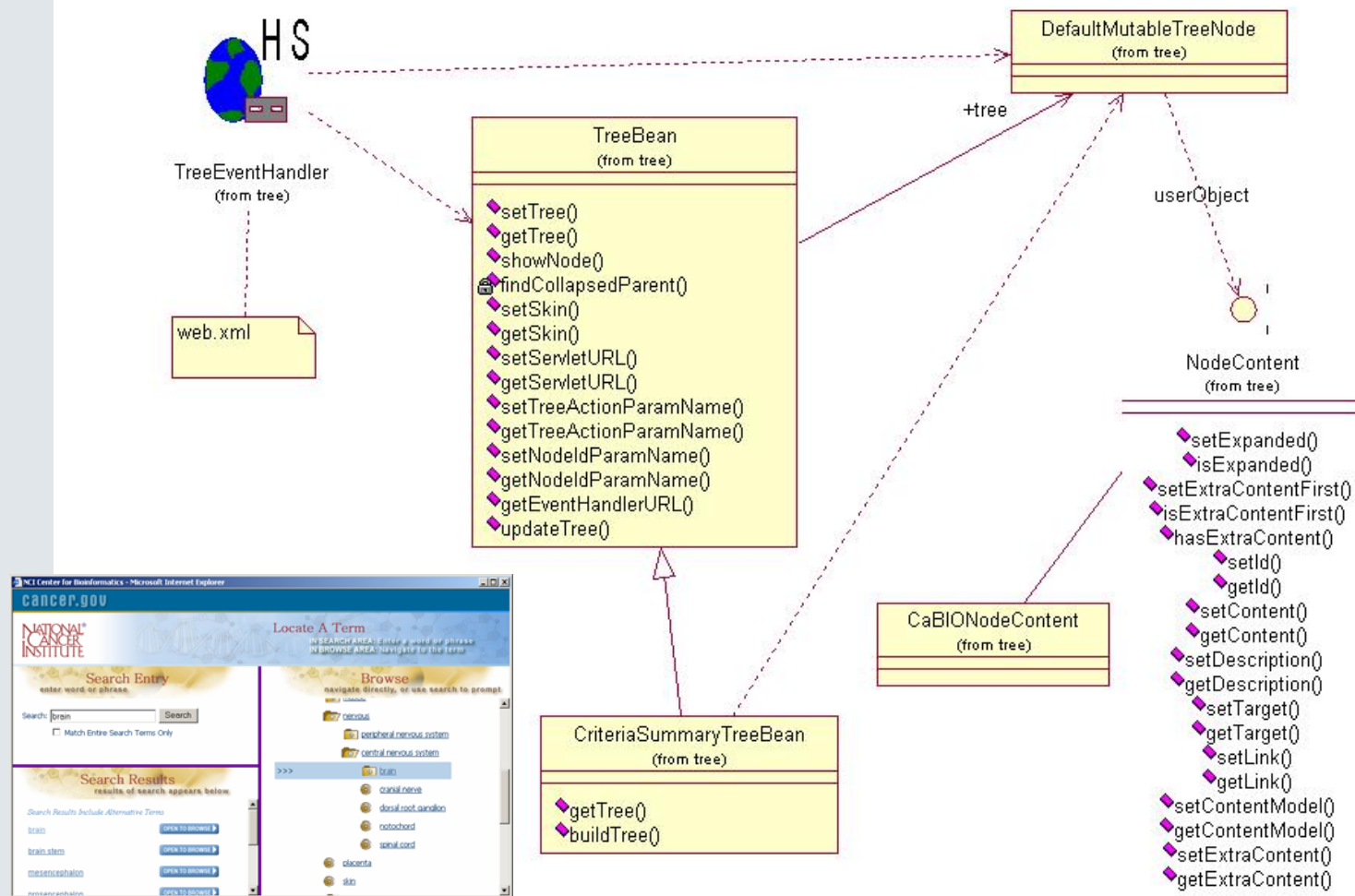
# High-level API



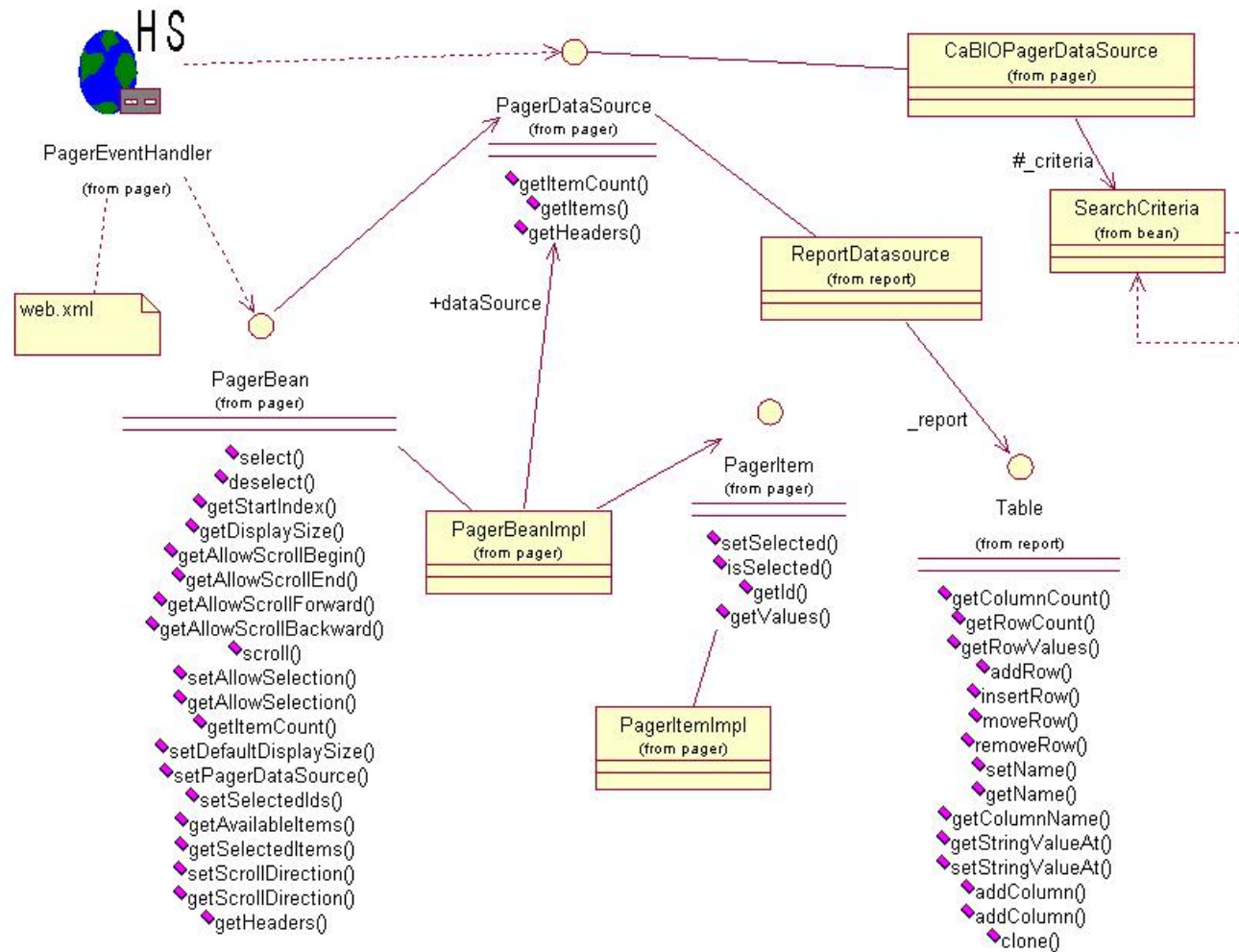
# Spreadsheet Parsing



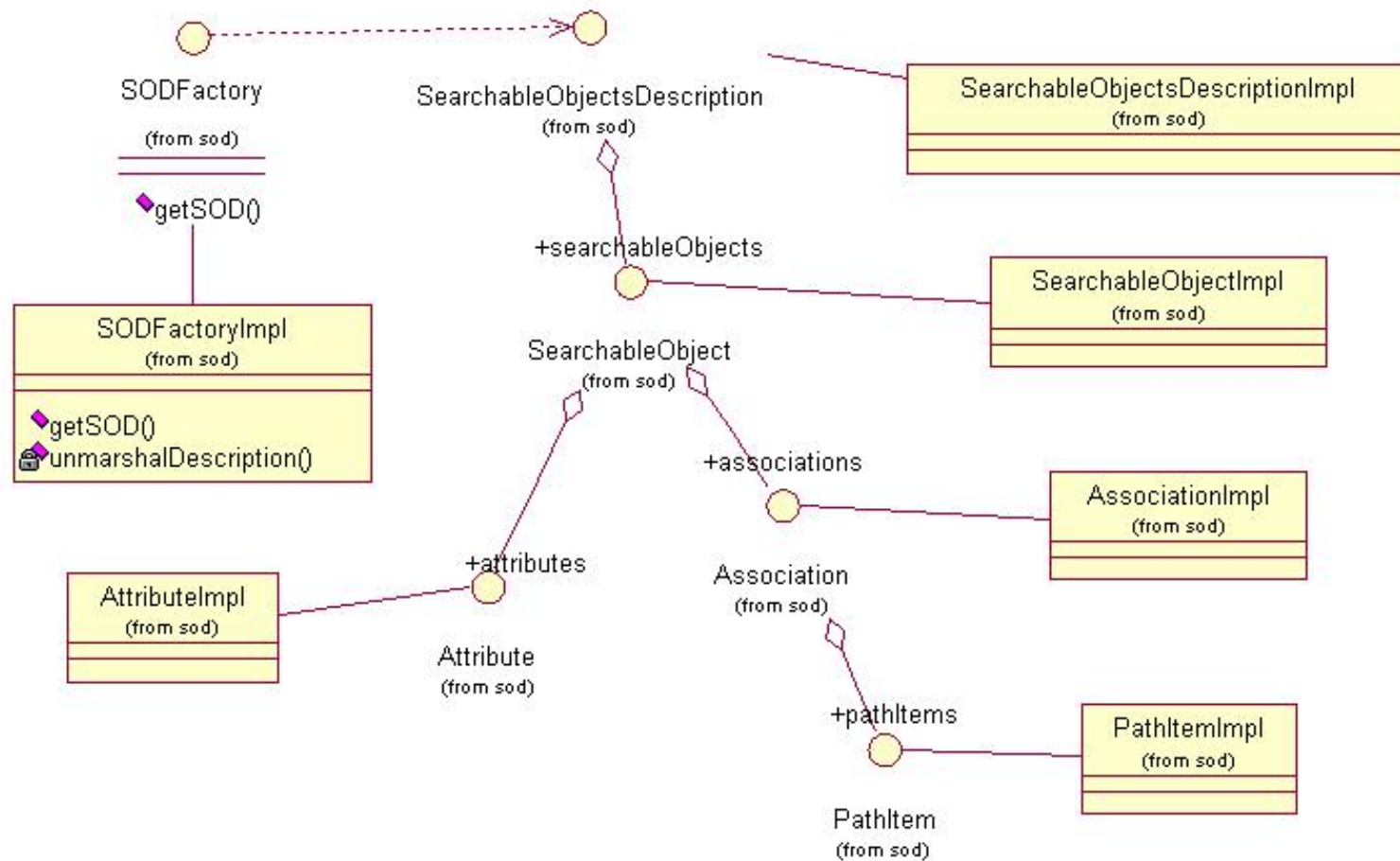
# Tree Manipulation



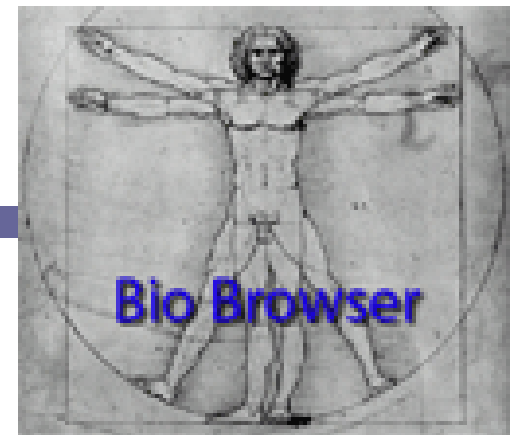
# Paging



# Metadata Layer



# Bio Browser



- ▶ **Bio Browser** is an excellent example of a desktop application built entirely on the caBIO HTTP interface.
- ▶ It's a Java Swing-based desktop application allows searching and browsing of the various objects served by the caBIO GetXML service.
- ▶ It produces a tree image of the underlying XML document, and also clickable pathway diagrams from the caBIO SVG data.
- ▶ Bio Browser is developed by Dr. Jonny Wray, a caBIO user who is developing a Bioinformatics course at UC Berkeley Extension.
- ▶ <http://www.jonnywray.com/java/index.html>

## caBIO Benefits

- ▶ **Abstraction Layer:** Provides an abstraction layer that allows developers to access bio medical information using a standardized tool set without concerns for implementation details
- ▶ **No Data Management:** Allows users to obtain information from a variety of data sources without data management concerns
- ▶ **Load Balancing:** Manages the display of large volumes of data to assist in load balancing
- ▶ **Complex Queries:** Provides an effective mechanism for performing complex queries that rely on diverse data sources
- ▶ **Provides Cross References:** Facilitates information sharing without managing linkages between multiple data sources

## caBIO Extensions

### ▶ Refactoring Effort

- Make all Object/Attribute/Methods consistent within the API.
- Consolidate common code.

### ▶ caBIO Perl API

- A complete Perl module based on caBIO web services that abstracts the SOAP marshalling for caBIO Perl users

### ▶ Extend Protein Objects

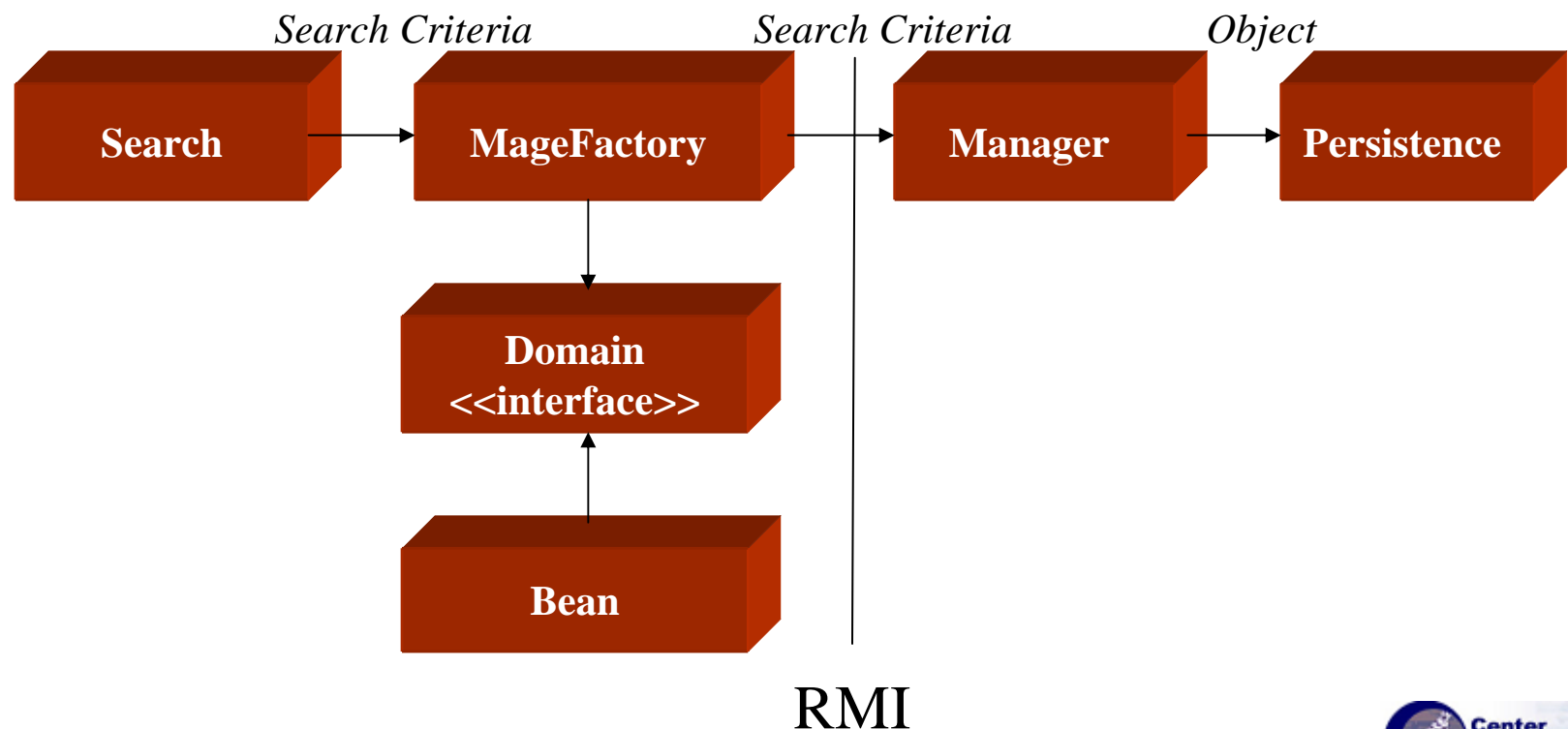
- Protein Portion of caBIO object model to accommodate UniProt and PDB Protein data with PIR Group

### ▶ Biogopher Navigator

- Based on graph theory and object search using Lucene

## MAGE-OM Architecture Details

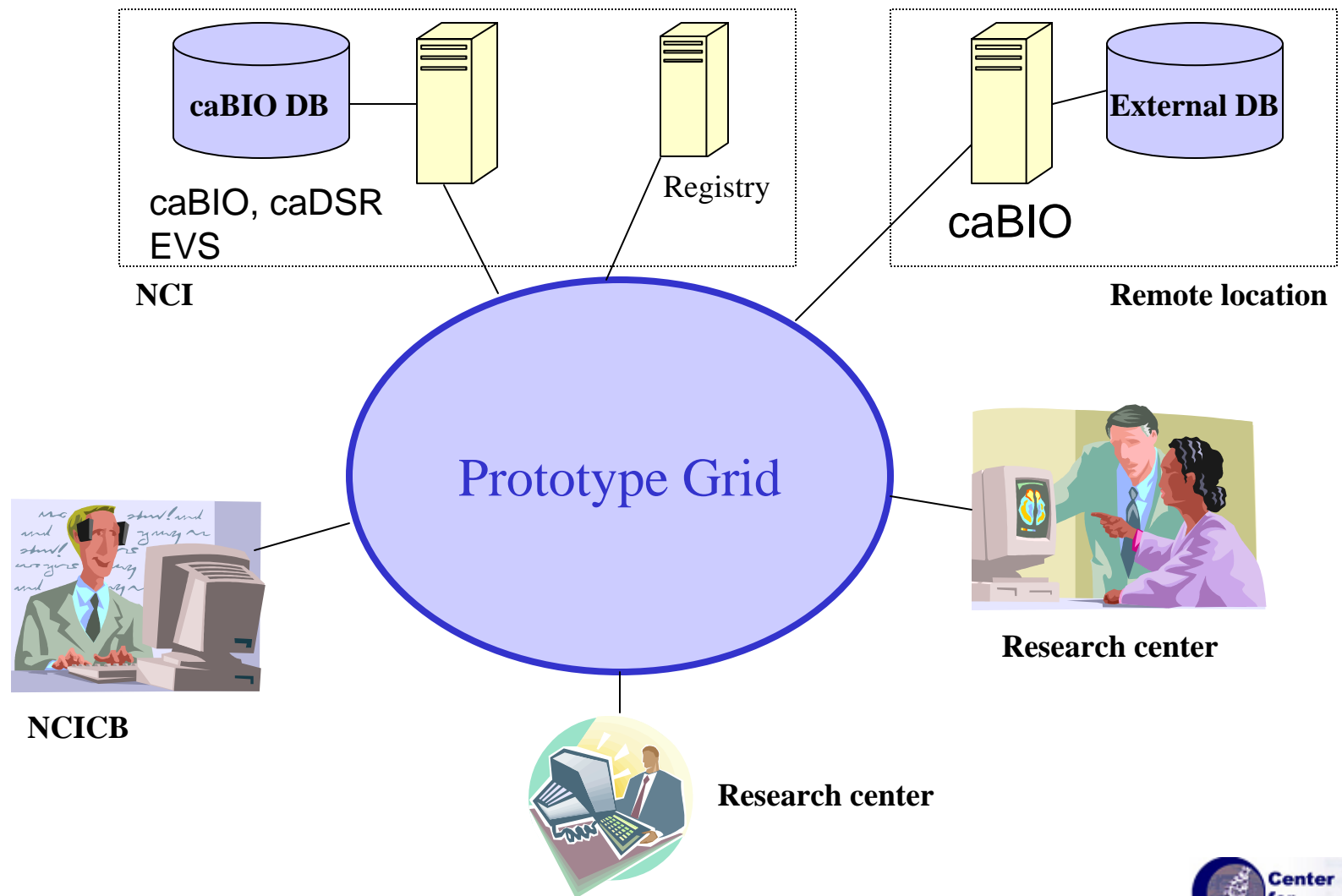
- ▶ The MAGE-OM is abstracted as Domain interfaces
- ▶ The Beans are implementations of the MAGE-OM interfaces but also provide caBIO methods (toXML)



## caBIO Kernel: Grid-Enabled Infrastructure

- ▶ Data collaboration across research centers.
- ▶ Share local data with other research centers.
- ▶ Data owners decide how much they want to share.
- ▶ Infrastructure upon which build applications.
- ▶ Infrastructure to instantiate object models.

# Grid-Enabled caBIO Prototype System



## Looking Around - Projects and Technologies

OGSA-DAI  
(Data grid)



(Data grid project)

Jena2  
(Semantics)



STORAGE  
RESOURCE  
BROKER

(Data grid application)

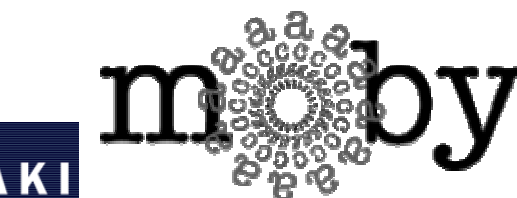


(Grid project)



(Grid infrastructure framework)

Web Services



(Web service registry for  
Bioinformatics )



(Data grid application)

JXTA  
(P2P technology)



(Grid project)

## Preliminary Grid Architecture

admin & configuration	GUI	Other applications
Grid		
Metadata	Metadata	Metadata
caBIO	Other Domain Object	Other data source

## Significant Use Case vs. Candidate Technologies

	Advertise	Discovery	Query	Obj. Mapping	Vocabulary
Web services	-UDDI -Extend UDDI. -Create a WS / server code.	-UDDI -Extend UDDI. -Create a WS / Server code.	- caBIO web services.		
Globus, OGSA-DAI, DQP	-Instantiate a grid service. -Registry new service. -Notification	-Indexing services (service data providers, data aggregators, grid service registry).	-Object model / caBIO java api. -Data bases (RDB, XML)		
SRB	- MCAP / Metadata service.	- MCAT / Metadata service.	- SRB server / FS, DB, Obj.		
Jena2	- Improve service description.	- Improve service discovery.	- Improve query when data mapped with EVS/caDSR.		-RWU Ontology languages. -Representation of semantic obj.
MCS / caDSR	- Metadata service	- Metadata service			
- OJB				Customize xml representation to model other DB.	

## caBIO is powered by!

- ▶ Dr. Ken Buetow
  - ▶ Dr. Peter Covitz
  - ▶ Dr. Carl Schaefer
  - ▶ Sharon Settnek
  - ▶ caBIO Team
  - ▶ caBIO Users!
- ▶ Open Source Technologies
    - Tomcat
    - Ant
    - Struts
    - OJB
    - Xalan
    - Xerces
    - XML-RPC
    - Jakarta Commons
    - POI
    - Junit
    - Zope
    - JAXB
    - SOAP
    - JDOM

